QoS as Middleware: Slot Based Bandwidth Brokering and Reservation[*]

(version 2.1)

Gary Hoo and William Johnston, Lawrence Berkeley National Laboratory

Ian Foster and Alain Roy, Argonne National Laboratory and University of Chicago

[*] This work is supported by the U. S. Dept. of Energy, Energy Research Division, Mathematical, Information, and Computational Sciences office, under contract DE-AC03-76SF00098 with the University of California and contract XXX with the University of Chicago. The authors may be contacted at: hoo@george.lbl.gov, wejohnston@lbl.gov, itf@mcs.anl.gov, roy@mcs.anl.gov. This document is report LBNL-xxxxx. For more information see www.itg.lbl.gov/Clipper .

The problem being addressed

Support for solving problems in grid computing environments that require aggregating many resources.

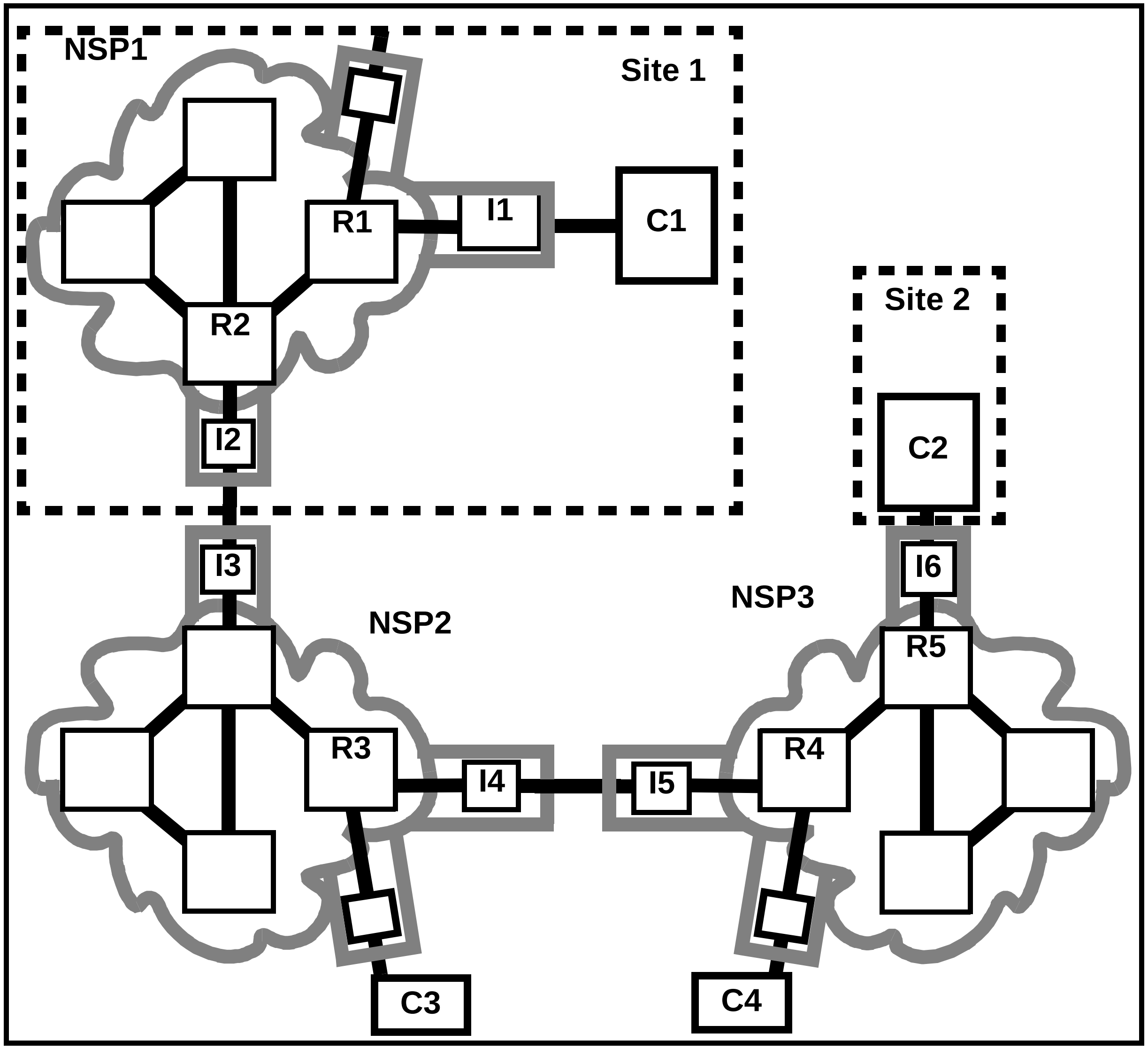
One aspect of this is the network bandwidth needed to connect other resources (e.g. CPUs, storage systems, instrument systems, etc.) that must act in concert.

This requires network quality of service that provides high priority bandwidth in shared networks that is:

- **reservable in advance**
- **managed by a system that can participate in negotiation**
- **not limited to low bandwidths**

The physical model

- sources and sinks (C1,2) at multiple sites
- multiple network service providers (NSP1,2,3) some of which might be site LANs
- NSP ingress points (I1,2,...) that provides
  - traffic conditioning
  - policy-based access control
  - accounting
- restriction points (R1,....) in the interiors of the networks that must be scheduled



Approach to diff-serv based bandwidth reservation

The network router services provide:

- ♦ **Priority service classes as the basic mechanism for giving some network traffic a higher queuing priority in router output buffers**
- ♦ **Shapers that can change a bursty flow into a smooth one of a specified bandwidth**
- ♦ **Membership in a priority class by having an edge router ("ingress" router) mark the packets destined for that class (and a higher queuing priority)**
  - **routers in the network interior provide differentiated service but generally do not mark packets**

Middleware services manage access to, and use of, the priority service class:

- ♦ **Service class admission control is provided by a reservation system separate from the router services**

- ♦ **The basic reservation unit is a *slot*:**
  - **some amount of bandwidth in some service class**
  - **a start and end time**

- ♦ **A slot manager keeps track of the allocated slots and the amount of bandwidth remaining to be allocated in the service class (and therefore effectively limits the amount of bandwidth that can be used in the service class)**

- ♦ **Requests for slots are mediated by an access control system**

- ♦ **Use of slots is recorded in an accounting record**

Further technical assumptions:

- ♦ **Fixed service level agreements exist between network service providers**

  (That is, adjacent NSPs will have a "business" agreement that says, e.g., NSP1 can automatically use a service class in NSP2 that provides up to X bits/sec of high priority traffic (e.g. at I3 and I5). Therefore, use of that SLA is just a slot reservation issue, not a policy issue.)

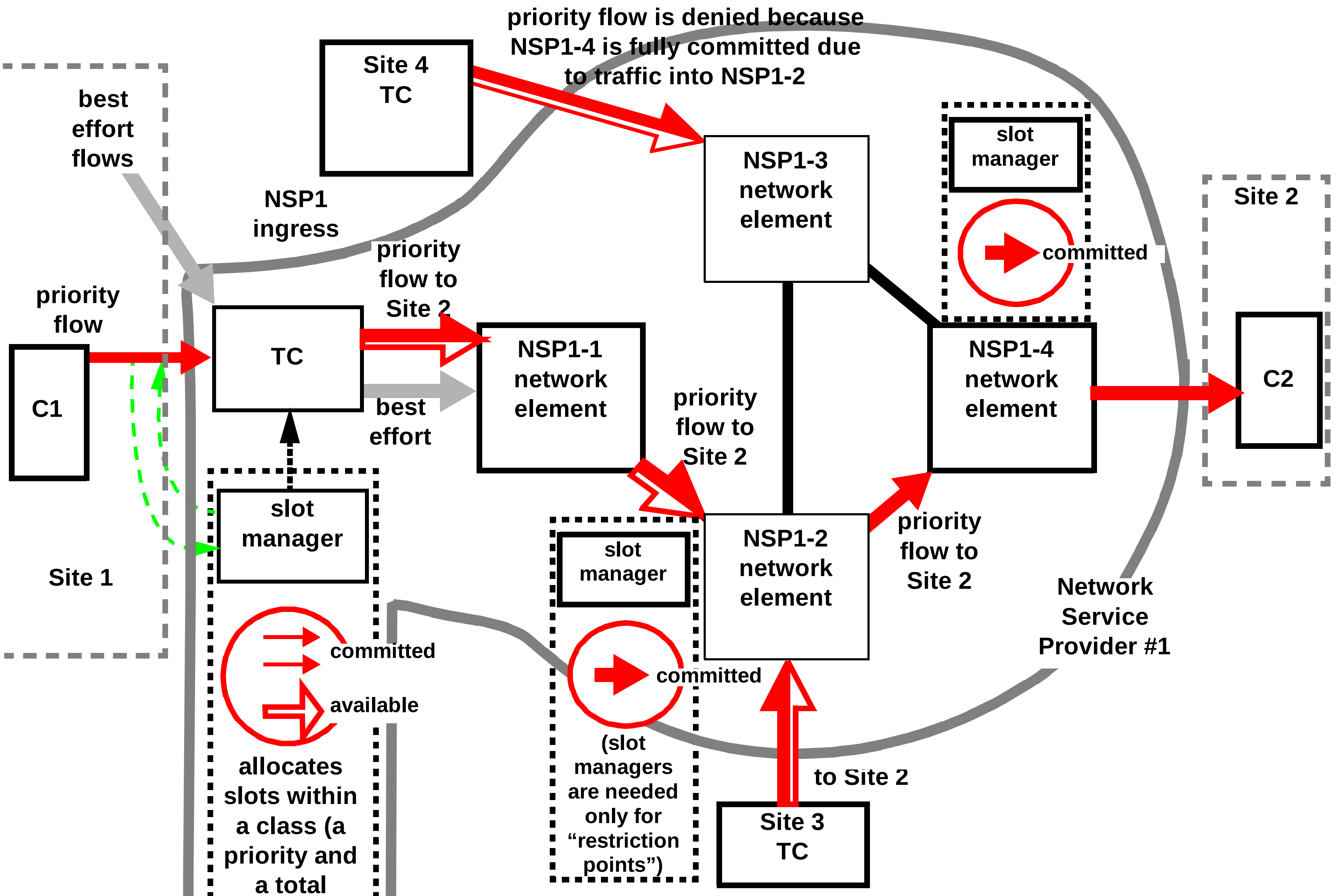
- ♦ **The following traffic conditioning mechanisms exist:**
  - **a traffic shaper at ingress nodes that can accept a flow that is, on average, within the bandwidth specification, but at the granularity of TCP writes, and modify that level of burstiness to whatever "smoothness" is required by the downstream network elements**
  - **a mechanism to place a flow into a special class (a marker)**
  - **a class based queuing mechanism (e.g. weighted fair queuing) that can provide differentiated queuing based on class membership throughout the network**

Design goals:

- Mechanism for advance, end-to-end reservation of bandwidth
- Mechanism for (simple) “negotiation”
  (I.e., a way to query for available priority bandwidth within some given range of bandwidth and time periods.)
- Mechanism for reservation path discovery end-to-end
- Network service provider has direct control over all resource utilization within its domain
- Policy-based access control for priority service
- Accounting for use of priority service
- Preemptive reservation cancellation mechanism with notification
- “Claiming” (of the reservation) must be light-weight and is done at flow start-up

- ♦ **Elements of the architecture:**
  - **network elements**
    1) **marker (part of ingress *traffic conditioner*)**
       - + **marks packets for a particular class**
       - + **controlled by slot manager**
       - + **probably in the ingress router**
    2) **shaper**
       - + **ensures that flow is within spec**
       - + **must accommodate 1-2 TCP windows worth of data per flow (because applications generally will only be able to regulate bandwidth at the socket level)**
       - + **maybe in the ingress router, maybe separate**
    3) **class-based queuer**
       - + **implements the queue discipline**
       - + **in all routers**
    4) **policers - anywhere deemed necessary by NSP**

- **slot manager**
  - **implements resource use policy (e.g. a resource is a premium traffic class, and the policy is that it will not be over-subscribed)**
  - **performs the basic reservation functions (allocating and de-allocating slots)**
  - **a resource can be anything that needs to be allocated to prevent congestion (e.g. a single router or switch, or a collection of these that are scheduled as a single resource because they represent a “restriction point” in the net)**



priority flow is denied because NSP1-4 is fully committed due to traffic into NSP1-2

best effort flows

Site 4 TC

NSP1 ingress

priority flow

C1

TC

priority flow to Site 2

best effort

NSP1-1 network element

slot manager

Site 1

committed

available

allocates slots within a class (a priority and a total bandwidth)

NSP1-3 network element

slot manager

committed

Site 2

C2

NSP1-4 network element

priority flow to Site 2

NSP1-2 network element

slot manager

committed

(slot managers are needed only for “restriction points”)

priority flow to Site 2

to Site 2

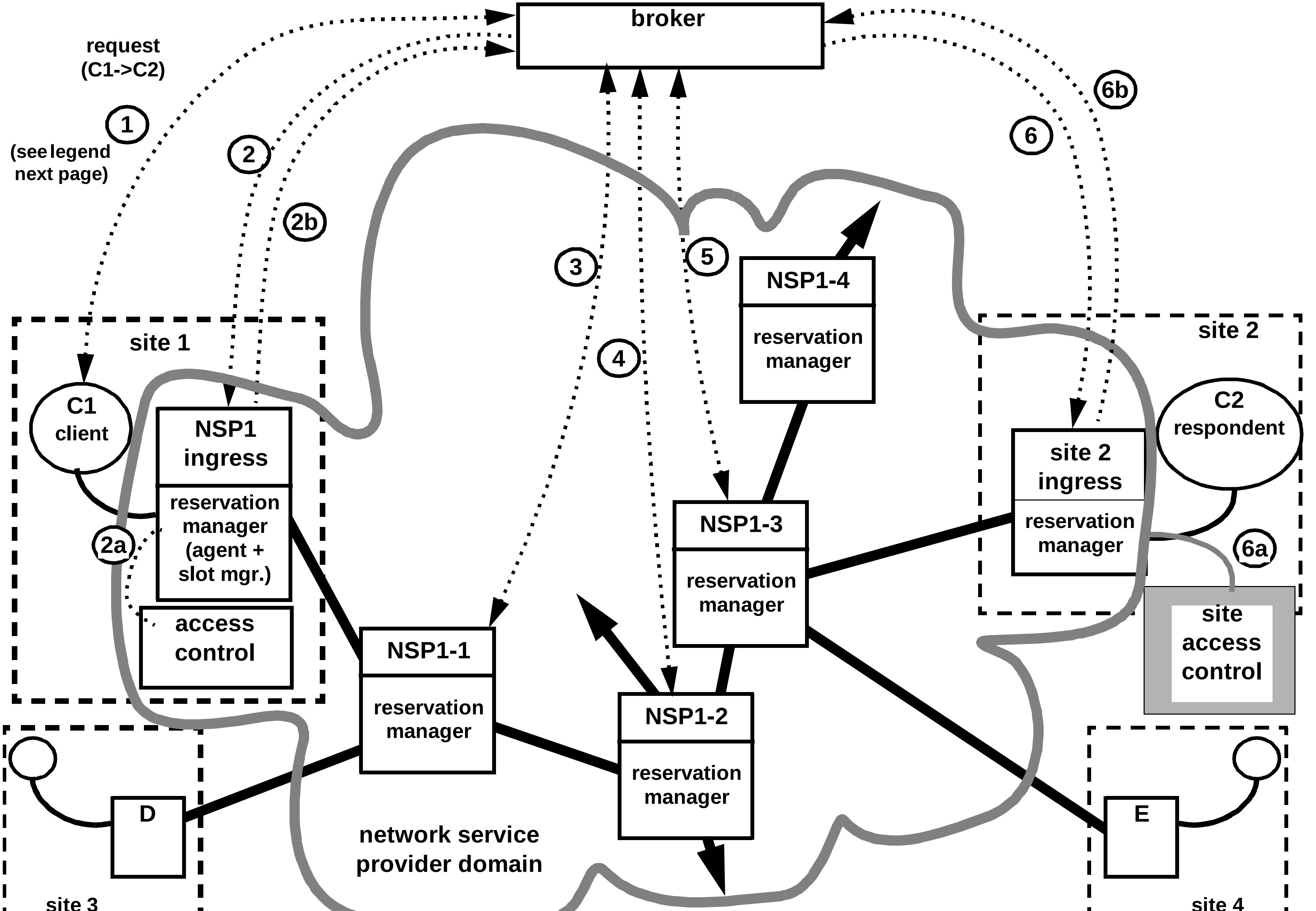
Site 3 TC

Network Service Provider #1

Slot Based Bandwidth Management (slot = time interval + bandwidth allocation)

- resource agents
  - brokerage interface to resource managers (slot managers)
  - provide simple negotiation
  - may represent multiple resources (i.e. multiple slot managers) - e.g., a single router, several routers, a whole domain of routers (i.e. a NSP), etc.
  - understands the topology of the network in order to provide identity of the “next” agent that must be contacted in order to form a “reservation path” through the network
  - under control of the resource owners (NSPs)
  - can invoke a policy-based access control engine to determine if it is willing to allocate a slot to a particular user
  - when at the ingress point, emits secure accounting record when reservation is claimed
  - probably returns a token representing the reservation

- broker
  - a general resource negotiator and aggregators
  - responsible for coordinating reservations on all of the resources (bandwidth on network paths, CPUs on multiple systems, etc.) required to accomplish a task
  - has to be able to query resources for available slots so that it may find a common time interval over all required resources
  - understands “path following” when “next agent to contact” is returned by resource agent



Reservation Request Phase

1) client C1 asks broker for premium bandwidth to C2

2) broker makes request of NSP ingress router agent with *client_id*

2a) ingress agent uses *client_id* to verify authority to use resource

2b) ingress agent returns to broker

- reservation token
  - *client_id*
  - authorization (as certified by ingress agent)
  - ingress resource reservation
- next-agent-to-contact

3) broker presents ingress reservation token to *next_agent* which:

- validates ingress authorization
- makes reservation against its allocation pool
- returns token and *next_agent* to broker

4) broker presents ingress reservation token to *next_agent*, etc.

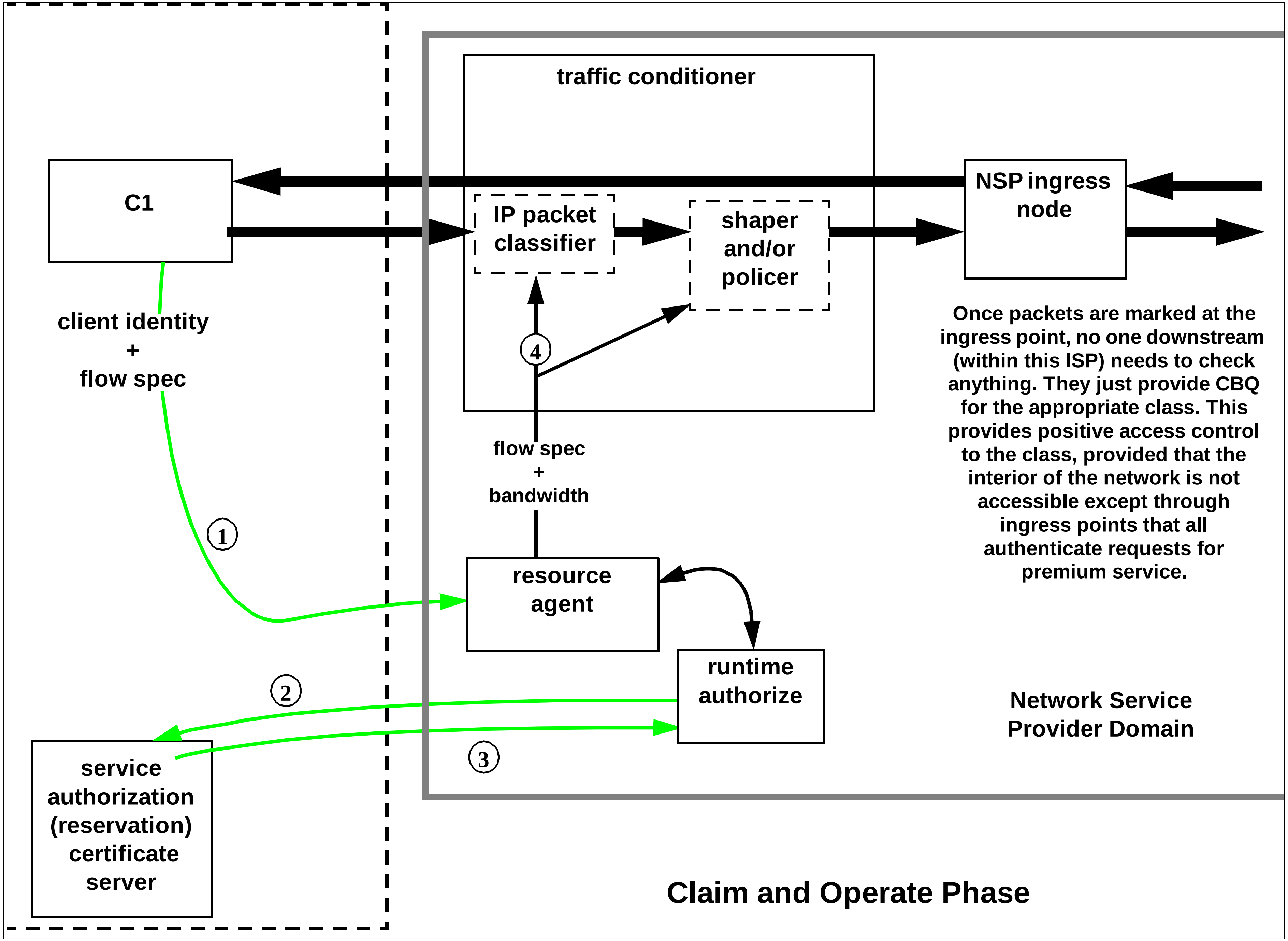
5) broker presents ingress reservation token to *next_agent*, etc.

(any individual reservation failure invalidates overall reservation)

6) broker presents ingress reservation token to *site_2* ingress agent

6a) *site_2* ingress agent requests authorization from the *site_2* access control system to use this resource (probably based on C2's proxy)

6b) *site_2* ingress agent returns reservation token to the broker with no *next_agent* and the reservation process is complete



Once packets are marked at the ingress point, no one downstream (within this ISP) needs to check anything. They just provide CBQ for the appropriate class. This provides positive access control to the class, provided that the interior of the network is not accessible except through ingress points that all authenticate requests for premium service.

Claim and Operate Phase

Status

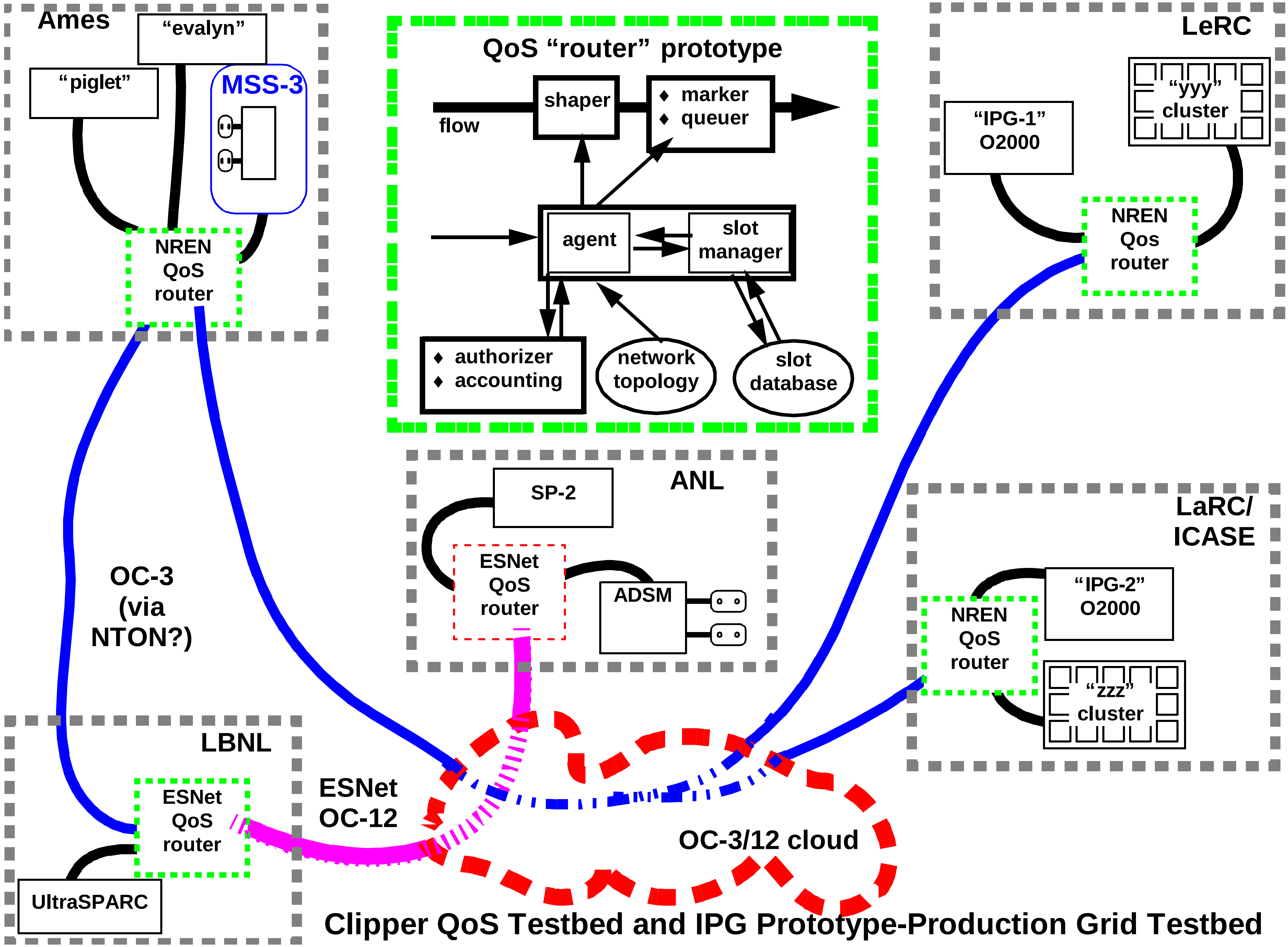
- slot manager
  - prototype implemented (Alain Roy and Gary Hoo)
- resource agent
  - interface designed (Gary Hoo, et al)
  - required functionality and near-term importance ():
    - (1) ability to identify "next agent to contact" - that is, given the destination address what is the next restriction point along the path
    - (3) respond with available slots "close" to the request
    - (3) manage several restriction points (several slot managers)
    - (2) ability to invoke policy engine to authorize user
    - (1) returns a token representing the reservation

- ♦ **broker**
  - **utility of Globus/DUROC needs to be determined**
  - **RSL extended for bandwidth slots (Foster, el at)**
- ♦ **network elements (*traffic conditioner*)**
  - **marker and class-based queuer**
    - **in Cisco routers at OC-3 - Van's original "bandwidth broker" illustrates use**
  - **shaper**
    - **may exist in routers at OC-3**
    - **for OC-12 use FreeBSD implementation (AltQ) (Tierney, Jin, et al)**
      - + **OC-12 NIC source agreement w/ Fore**
      - + **40 MBy/sec PC configuration identified**

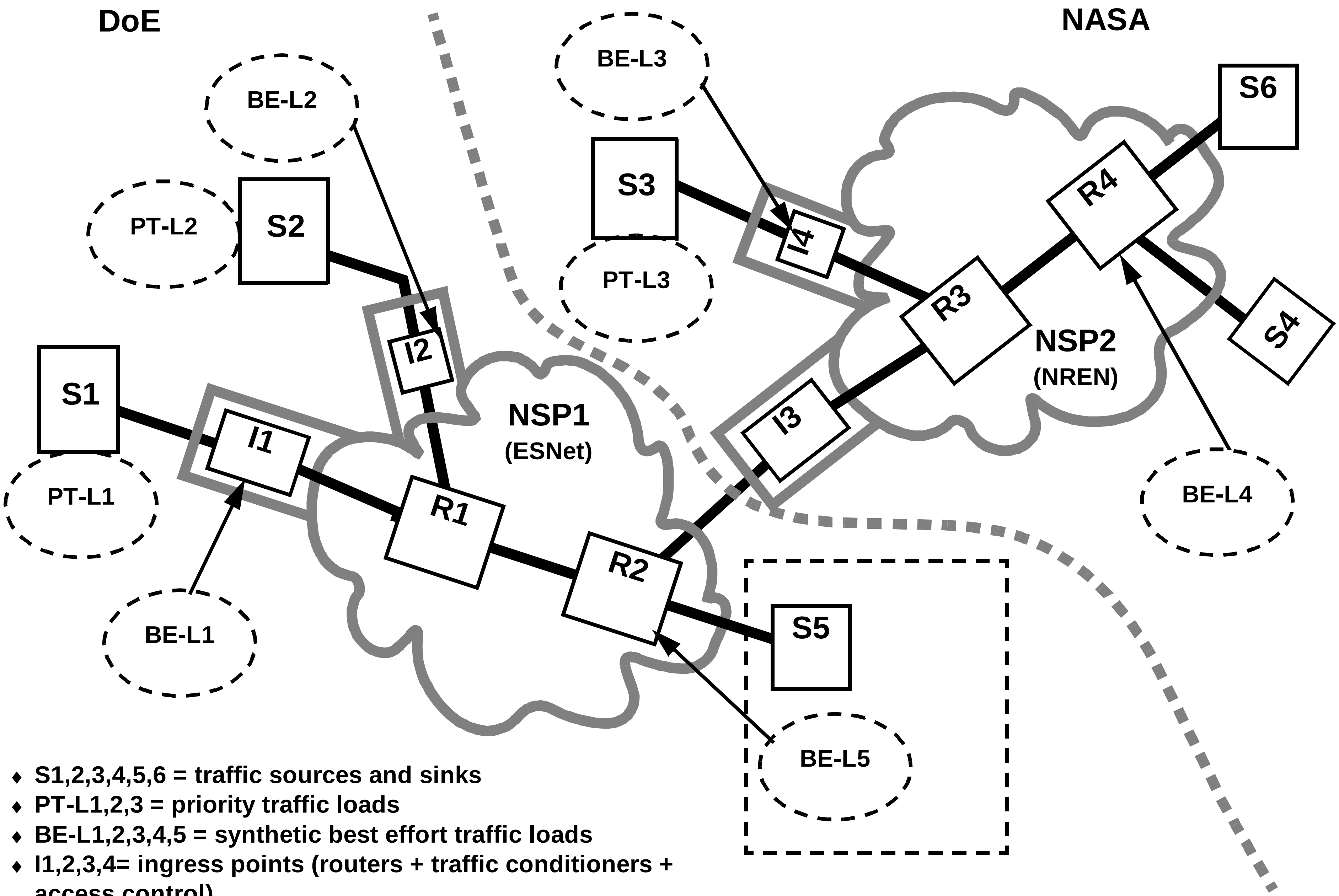
- **QoS testbed**

  For discussion (see next page):

  - **Foster/Cisco equipment grant**
  - **Cisco COPS/RSVP proposal (see http://www-itg.lbl.gov/Clipper/private)**
  - **Ames - LBNL OC-3 connectivity (via NTON?)**
  - **Commercial ATM service issues that would impact effectiveness of QoS routers at DOE Labs / NASA Centers**
  - **Network engineering issues**



Clipper QoS Testbed and IPG Prototype-Production Grid Testbed



- S1,2,3,4,5,6 = traffic sources and sinks
- PT-L1,2,3 = priority traffic loads
- BE-L1,2,3,4,5 = synthetic best effort traffic loads
- I1,2,3,4= ingress points (routers + traffic conditioners + access control)
- R1,2,3,4 = restriction points (routers)
- NSP1,2 = network domains

Proposed LBNL-ANL-ESNet-NREN QoS testbed